Aff-Wild: Valence and Arousal 'in-the-wild' Challenge

Stefanos Zafeiriou[⋆,3] Dimitrios Kollias[⋆ *] Mihalis A. Nicolaou[†]
Athanasios Papaioannou[⋆] Guoying Zhao[3]
Irene Kotsia[1,2]
[⋆]Department of Computing, Imperial College London, UK
[†]Department of Computing, Goldsmiths, University of London, UK
[3]Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[1] School of Science and Technology, International Hellenic University, Greece
[2] Department of Computer Science, Middlesex University, UK
[⋆]{s.zafeiriou, dimitrios.kollias15}@imperial.ac.uk, [†]m.nicolaou@gold.ac.uk

Abstract

The Affect-in-the-Wild (Aff-Wild) Challenge proposes a new comprehensive benchmark for assessing the performance of facial affect/behaviour analysis/understanding 'in-the-wild'. The Aff-wild benchmark contains about 300 videos (over 2,000 minutes of data) annotated with regards to valence and arousal, all captured 'in-the-wild' (the main source being Youtube videos). The paper presents the database description, the experimental set up, the baseline method used for the challenge and finally the summary of the performance of the different methods submitted to the Affect-in-the-Wild Challenge for Valence and Arousal estimation. The challenge demonstrates that meticulously designed deep neural networks can achieve very good performance when trained with in-the-wild data.

1. Introduction

Behavioral modeling and analysis constitute a crucial aspect of Human Computer Interaction. Emotion recognition is a key issue, dealing with multimodal patterns, such as facial expressions, head pose, hand and body gestures, linguistic and paralinguistic acoustic cues, as well as physiological data. However, generating machines which are able to recognize human emotions is a difficult problem, because the emotion patterns are complex, time-varying, user and context dependent, especially when considering uncontrolled environments, i.e., 'in-the-wild'.

Current research in automatic analysis of facial affect aims at developing systems, such as robots and virtual humans, that will interact with humans in a naturalistic way under real-world settings. To this end, such systems should automatically sense and interpret facial signals relevant to emotions, appraisals and intentions. Furthermore, since real-world settings entail uncontrolled conditions, where subjects operate in a diversity of contexts and environments, systems that perform automatic human behaviour analysis should be robust to video recording conditions, the diversity of contexts and the timing of display.

The past twenty years research in automatic analysis of facial behaviour was mainly limited to posed behavior captured in highly controlled recording conditions [29, 35, 33, 24]. Some representative datasets, which are still used in many recent works [18], include the Cohn-Kanade database [33, 24], the MMI database [29, 35], the Multi-PIE database [17] and the BU-3D and BU-4D databases [41, 40]. Nevertheless, it is now accepted by the community that the facial expressions of naturalistic behaviour could be radically different from the posed ones [9, 32, 44]. Hence, efforts have been made in order to collect subjects displaying naturalistic behaviour. Examples include the recently collected EmoPain [4] and UNBC-McMaster [27] for analysis of pain, the RU-FACS database consisting of subjects participating in a false opinion scenario [5] and the SEMAINE [27] corpus which contains recordings of subjects interacting with a Sensitive Artificial Listener (SAL) under controlled conditions. All the above databases have been captured in well-controlled recording conditions and mainly under a strictly defined scenario (e.g., pain estimation).

Representing human emotions has been a basic topic of research in psychology. The most frequently used emotion representation is the categorical one, including the seven basic categories, i.e., Anger, Disgust, Fear, Happiness, Sadness, Surprise and Neutral [13][10]. It is, however, the dimensional emotion representation [39, 31] which is more

*The first two authors contributed equally in the paper.

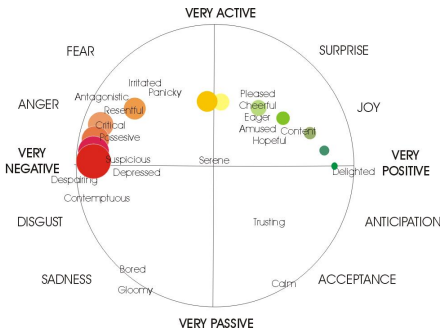


Figure 1: The 2-D Emotion Wheel

appropriate to represent subtle, i.e., not only extreme, emotions appearing in everyday human computer interactions. The 2-D Valence and Arousal Space is the most usual dimensional emotion representation. Figure 1 shows the 2-D Emotion Wheel [30], with valence ranging from very positive to very negative and arousal ranging from very active to very passive.

There are various signs of a humans emotions, such as facial expressions, gestures, paralinguistic speech features and physiological measurements. In the Challenge we focus on facial affect as measured by valence and arousal annotations. In particular, we make a considerable effort to go beyond the current practices in facial behaviour analysis and collect and annotate the first large scale in-the-wild database of facial affect [1]. To achieve this, we capitalise on the abundance of data available in video-sharing websites, such as YouTube [42] [2], and select videos that display the affective behavior of people, for example videos that display the behaviour of people when watching a trailer, a movie, a disturbing clip or reactions to pranks etc. To this end we have collected 298 videos displaying reactions of 200 subjects. To the best of our knowledge this is the largest database containing videos of facial behaviour "in-the-wild". For a recent survey on facial behaviour analysis in-the-wild with an emphasis on deep learning methodologies the interested reader may refer to [43]. This database has been annotated by 6-8 lay experts with regards to two continuous emotion dimensions, i.e. valence, which records how positive or negative an emotion is, and arousal which measures the power of the activation of the emotion.

In the rest of the paper, we first describe the generated Aff-Wild database (Section 2), afterwards we describe the annotation procedure (Section 3) and then we present the results of the challenge (Section 3). Subsequently, in Section 4.1 we make a reference to all methods and respective papers submitted to the Challenge, summarize the obtained results in valence and arousal estimation and declare the winning method.

2. The Aff-Wild Database

We collected a database consisting of 298 videos, with a total length of more than 30 hours. The aim was to collect spontaneous facial behaviors under arbitrary recording conditions. To this end, the videos were collected using the Youtube video sharing web-site. The keyword used to retrieve the videos was "reaction"; the resulting videos display subjects reacting to a variety of stimuli (e.g., tasting something hot or disgusting). Examples include subjects reacting to an unexpected plot twist of a movie or series, a trailer of a highly anticipated movie, etc. The subjects display both positive and negative emotions (or combinations of them). In other cases, subjects display emotions while performing an activity (e.g., riding a rolling coaster). In some videos, subjects react on a practical joke, or on positive surprises (e.g., a gift). Most of the videos were in YUV 4:2:0 format, with some of them being in AVI format; all have been annotated in terms of valence and arousal. Six to eight subjects have annotated the videos following a methodology similar to the one proposed in [11]. That is, an on-line annotation procedure was used, according to which annotators were watching each video and provided their annotations through a joystick. Valence and arousal ranged continuously in $[-1, +1]$. We have annotated all subjects that are present in a video. In total we have 200 subjects, with 130 of them being male and 70 of them female. Figures 2 and 3 demonstrate some frames of the Aff-Wild database.

In Figures 4, 5 we present two characteristic examples of facial images, cropped from two different videos, with their respective video frame number and the valence and arousal annotation for each of them. We also present a visual representation of these values on the 2-D emotion space, showing the change of the reactions/behavior of the person among these time instances of the video. Time evolution is indicated, by using a larger size for the more recent frames and a smaller size for the older ones.

It can be verified that annotations correspond well to the facial expression displaying in the video frames. It should, however, be added that it is often difficult to say which is the true emotional state of the acting person from a static frame. That is why, working with many annotators and selecting

[1] Currently, there are many challenges in behaviour analysis, including the series of AVEC [37, 36, 34] challenges. Nevertheless, AVEC uses only data captured in controlled conditions and under very specific scenarios. The only challenge that uses 'in-the-wild' data is the series of [15, 14, 16]. Nevertheless, the samples come from movies and the annotation is limited to the universal expressions.

[2] The collection has been conducted under the scrutiny and approval of Imperial College Ethical Committee (ICREC). The majority of the chosen videos were under Creative Commons License (CCL). For those videos that were not under CCL, we have contacted the person who created them and asked for their approval to be used in this research.

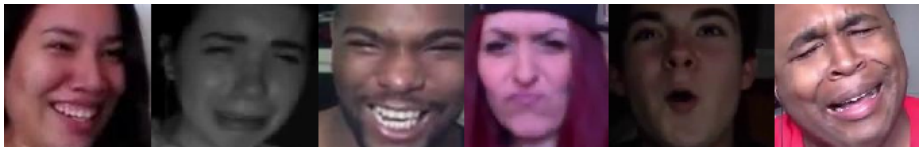


Figure 2: Some representative frames from the Aff-Wild database.

Figure 3: Some challenging frames from the Aff-Wild database.

the ones that are more consistent between them is necessary to get more accurate annotation of the underlying emotion.

Table 1: Number of Subjects in the Aff-Wild Database

Database	no of males	no of females
Train	106	48
Test	24	22

Table 2: Attributes of the Aff-Wild Database

Attribute	Description
Length of videos	0.10-14.47 min
No of annotators	6-8
Total no of videos	252(train)+46(test) = 298
Video format	AVI , MP4

3. Annotation and data processing

3.1. Annotation tool

For data annotation, we developed our own application which was similar to others like Feeltrace [11] and Gtrace [12]. In our application we used a setting with one time-continuous annotation for each affective dimension, like in Gtrace. We did not want to judge valence and arousal at the same time, like in Feeltrace, because it would be too cognitively demanding to reach a high quality on both. The user at first selects if (s)he wants to annotate valence or arousal. Then, the interface of our application asks the user to log in using an identifier, his/her name, and to select an appropriate joystick. After that, the screen is split into two parts: a scrolling list of all videos is given on the left side and on the right side there is a scrolling list of all annotated videos. After one selects a video to annotate, a screen appears that shows the video and a slider of values ranging in $[-1, 1]$. Then the video can be annotated by moving the joystick either up or down. At the same time our application samples the annotations at a variable time rate. Figure 6 shows the graphical interface of our tool when annotating valence (the tool for arousal is similar).

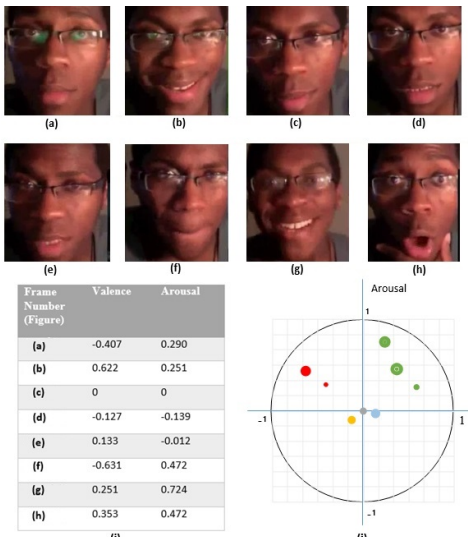
3.2. Annotation guidelines

Each annotator was instructed orally and received instructions through a multi page document, explaining in detail the procedure to follow for the annotation task. This document included a short list of some well identified emotional cues for both arousal and valence, in order to provide a common introduction on emotions to the annotators, even though they were rather instructed to use their own feeling for the annotation task [3]. Before starting the annotation of the data, each annotator watched the whole video so as to know what to expect regarding all emotions being depicted in the video.

3.3. Data pre-processing

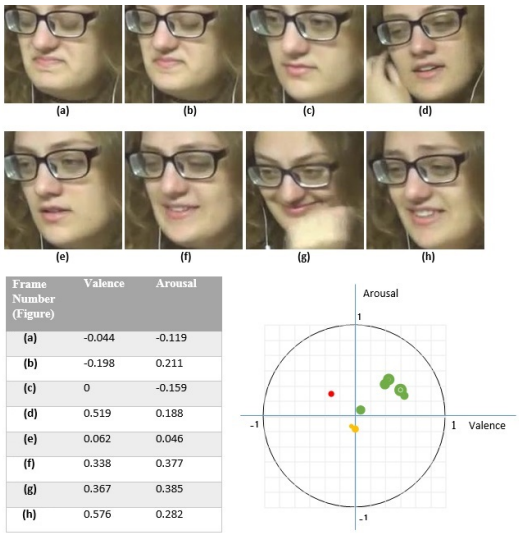
VirtualDub [22] was used in order to trim the raw YouTube videos, mainly at the start and the end of them, so as to remove useless content (e.g., an advertisement). Then another pre-processing step was applied in order to locate the faces in all frames of the videos. In more detail,

[3] All annotators were computer scientists who were working on face analysis problems and all had a working understanding of facial expressions.



Frame Number (Figure)	Valence	Arousal
(a)	-0.407	0.290
(b)	0.622	0.251
(c)	0	0
(d)	-0.127	-0.139
(e)	0.133	-0.012
(f)	-0.631	0.472
(g)	0.251	0.724
(h)	0.353	0.472

Figure 4: Annotated Facial Expressions (Person A)



Frame Number (Figure)	Valence	Arousal
(a)	-0.044	-0.119
(b)	-0.198	0.211
(c)	0	-0.159
(d)	0.519	0.188
(e)	0.062	0.046
(f)	0.338	0.377
(g)	0.367	0.385
(h)	0.576	0.282

Figure 5: Annotated Facial Expressions (Person B)

we extracted a total of 1,180,000 frames using the Menpo software [2]. From each frame, we detected the faces using the method described in [26]. We also developed a matching process between the annotation time stamps and the cropped faces time instances. In particular, for each frame time instance, we searched for the nearest neighbor, in the annotation time stamp sequence and then linked the latter valence and arousal annotation values to the corresponding frame time stamp. In cases where we had two annotation timestamps with same time distance from a frame, we computed the average of those two timestamps and attributed this value to that frame. Finally, we extracted facial landmarks for all frames using the best performing method in [8].

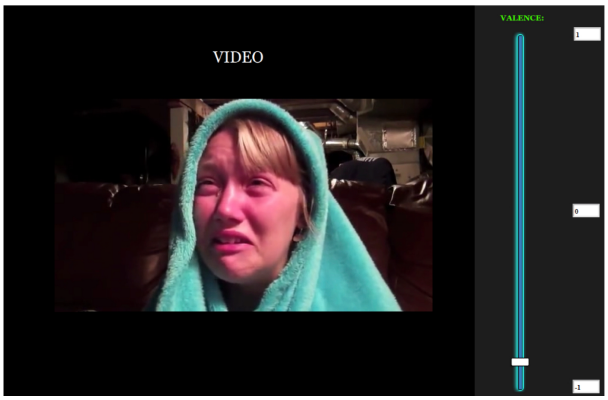


Figure 6: The GUI of the annotation tool when annotating valence (the GUI for arousal is exactly the same).

3.4. Annotation Post-processing

We further extended our annotation tool so that it plays a specific video and at the same time plots the corresponding valence and arousal annotated values. After the annotation was completed, every annotator was asked to watch the videos again and verify that the corresponding annotations were indeed in accordance with the annotator's perception of the emotional state expressed in the videos. In this way, a further validation of annotations was achieved. After the annotations have been validated, we computed, for every video, cross-correlations between all annotators. Furthermore, we computed the correlation between the annotation and the tracked facial landmarks. As a consequence, we ranked the annotators' correlation for each video. Two more experts then watched all videos and, for every video, selected the most correlated best annotations (between 2 to 4 annotations). We then computed the mean of these annotations; in other words we selected the mean of annotations that were mostly correlated and were given good evaluation by the experts. Figure 7 shows a small part of a video (2000 frames) and the 4 most highly correlated annotations for valence.

Figure 8 provides a histogram for the annotated values for valence and arousal in the generated database. As it can be observed the annotation is currently biased towards positive valence and arousal (something we aim at addressing in future runs of the challenge).

4. Experiments with a Baseline

Due to the fact that it was the first time that estimation of valence and arousal was attempted in in-the-wild videos the standards approaches, e.g. using Support Vector Regression (SVR) etc. on image features [28], resulted in very poor performance. Hence, we implemented a deep neural network approach as the baseline. Currently deep neural networks provide the state-of-the-art in many tasks in computer vision, speech analysis and natural language

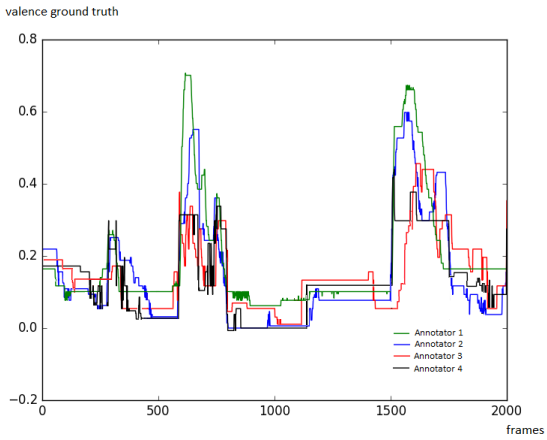


Figure 7: The 4 most highly correlated valence annotations over a part of a video

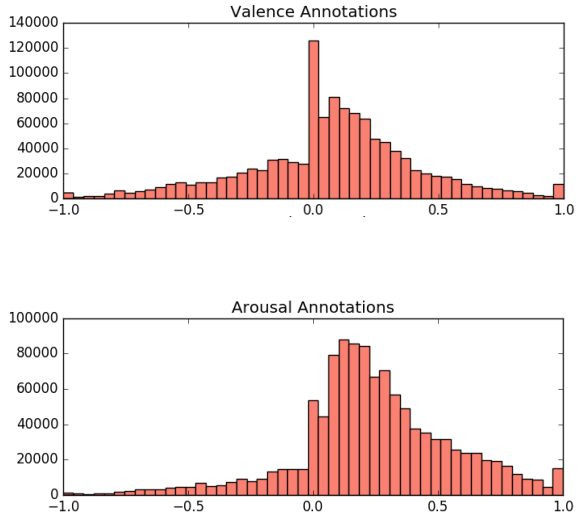


Figure 8: Histogram of Annotations

processing that require learning from massive data. They have also achieved good performances in emotion recognition challenges and contests [15]. Deep Convolutional Neural Networks (CNNs) [20] [21] include convolutional layers with feature maps composed of neurons with local receptive fields, shared weights and pooling layers, which are able to automatically extract non-linear features. The baseline architecture was based on the structure of the CNN-M [7] network. We used the pre-trained on the FaceValue dataset [3] CNN-M network as starting structure and performed transfer learning of its convolutional and pooling parts on our designed network. In particular, we used two fully connected layers, the second being the output layer providing the valence and arousal predictions. We either froze the CNN part of the network and performed fine-tuning of the weights of the fully connected layers, or performed fine-tuning of the weights of the whole network. The exact structure of the network is shown in Table 3. Note that the activation function in the convolutional and batch normalisation layers is the ReLu one; this is also the case in the first fully connected one. The activation function of the second fully connected layer is linear. It should be also mentioned that the utilized deep learning architecture has been implemented on the TensorFlow platform [1]. For the pre-trained network we took the one in MatConvNet [38] and transformed it into a format recognisable by TensorFlow. Also note that we follow the TensorFlow's platform notation for the sizes of all parameters of the convolutional and pooling layers.

Table 3: Baseline Architecture based on CNN-M showing the sizes for the parameters of the convolutional and pooling layers and the no of hidden units in the fully connected ones

Layer	filter	ksize	stride	padding	no of units
conv 1	[7, 7, 3, 96]		[1, 2, 2, 1]	'VALID'	
batch norm					
max pooling		[1, 3, 3, 1]	[1, 2, 2, 1]	'VALID'	
conv 2	[5, 5, 96, 256]		[1, 2, 2, 1]	'SAME'	
batch norm					
max pooling		[1, 3, 3, 1]	[1, 2, 2, 1]	'SAME'	
conv 3	[3, 3, 256, 512]		[1, 1, 1, 1]	'SAME'	
batch norm					
conv 4	[3, 3, 512, 512]		[1, 1, 1, 1]	'SAME'	
batch norm					
conv 5	[3, 3, 512, 512]		[1, 1, 1, 1]	'SAME'	
batch norm					
max pooling		[1, 2, 2, 1]	[1, 2, 2, 1]	'SAME'	
fc 1					4096
fc 2					2

For training the network (in mini batches) we used the Adam optimizer algorithm. The Mean Squared Error (MSE) was used as the error/cost function. The hyper-parameters being used were: the batch size which was 80, the constant learning rate being 0.001 and the number of hidden units in the first fully connected layer which was 4096. We also used biases in the fully connected layers. The weights of the fully connected layers were initialised from a Truncated Normal distribution with a zero mean and variance equal to 0.1 and the biases were initialised to 1. Training was performed on a single GeForce GTX TITAN X GPU and the training time was about 4-5 days. No data augmentation techniques were used, because the database was already large enough. For training the baseline the training data have been split to two sets. One was used for training the network and the other for validation.

Table 4 summarizes the obtained MSE and Concordance Correlation Coefficient (CCC) Values over the whole generated database by our baseline network. It should be mentioned that the CCC is defined as

$$\rho_c = \frac{2 s_{x_y}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \quad (1)$$

where s_x and s_y denote the variance of the predicted and

ground truth values respectively, $\bar{x}$ and $\bar{y}$ are the corresponding mean values and s_{x_y} is the respective covariance value. From the results we deduced that the task is very challenging and requires meticulously designed deep learning architectures in order to be tackled.

Table 4: Concordance (CCC) and Mean Squared Error (MSE) evaluation of valence & arousal predictions provided by the CNN M baseline architecture

	CCC		MSE	
	Valence	Arousal	Valence	Arousal
CNN M	0.15	0.10	0.13	0.14

4.1. The AFF-Wild Challenge

The training data (i.e., videos and annotations) of AFF-wild challenge were made publicly available on the 30th of January 2017. The test videos (without annotations) were made available around 22nd of March. The participants could submit an entry to the challenge until 2nd of April.

Ten different research groups were initialy interested in the Aff-Wild challenge. These groups downloaded the datasets and enquired about different issues of the challenge. Six of them made experimentation and submitted their results to the Workshop portal. Based on the performance they obtained on the test data, three of them finally submitted a paper to the workshop. These are briefly reported below, while Table 5 compares the derived results (in terms of CCC and MSE) by all three methods.

Table 5: Concordance (CCC) and Mean Squared Error (MSE) of valence & arousal predictions provided by the 3 methods

Methods	CCC		MSE	
	Valence	Arousal	Valence	Arousal
MM-Net	0.196	0.214	0.134	**0.088**
FATAUVA-Net	**0.396**	**0.282**	**0.123**	0.095
DRC-Net	0.042	0.291	0.161	0.094

In [23] (Method MM-Net), a variation of the deep convolutional residual neural network is first presented for affective level estimation of facial expressions . Then multiple memory networks are used to model temporal relations between the video frames. Finally, ensemble models are used to combine the predictions of the multiple memory networks, showing that the latter steps improve the initially obtained performance, as far as MSE is concerned, by more than 10%.

In [6] (Method FATAUVA-Net), a deep learning framework is presented, in which a core layer, an attribute layer, an AU layer and a V-A layer are trained sequentially. The facial part-based response is firstly learnt through attribute recognition Convolutional Neural Networks, and then these layers are applied to supervise the learning of AUs. Finally, AUs are employed as mid-level representations to estimate the intensity of valence and arousal.

In [25] (Method DRC-Net), three neural network-based methods are presented and compared, which are based on Inception-ResNet modules redesigned specifically for the task of facial affect estimation. These methods are: Shallow Inception-ResNet, Deep Inception-ResNet, and Inception-ResNet with LSTMs. Facial features are extracted in different scales and simultaneously both the valence and arousal are estimated in each frame. Best results in the experiments are obtained by the Deep Inception-ResNet method.

All participants applied deep learning methods to the problem. The winning method of our Challenge is FATAUVA-Net Method, since it achieved the best results in the majority of the metrics used.

4.2. Additional Experiments

As organizers we could not participate in the Aff-W challenge. Nevertheless, we believe that it would be interesting to report results with a method that we were developing during the course of the challenge. This method was based on an end-to-end architecture composed of CNNs and Recurrent Neural Networks (CNN-RNN) which was trained and tested using the Aff-Wild benchmark. In particular, we used different pre-trained CNN and performed transfer learning and fine-tuning for designing the CNN part of the network. By including an RNN part and retraining the resulting end-to-end architecture we obtained our best results for valence and arousal estimation. More details about this method are reported in [19] (Method VA-CRNN). As can be seen, we have been able to achieve better results than those achieved by the participants of the challenge. These results are shown in Table 6. The above results demonstrate that even though the task is very challenging an elaborate deep learning approach could achieve very good results (around 0.57 CCC for valence and 0.43 for arousal)

Table 6: Concordance (CCC) and Mean Squared Error (MSE) of valence & arousal predictions provided by our method

Method	CCC		MSE	
	Valence	Arousal	Valence	Arousal
VA-CRNN	0.57	0.43	0.08	0.06

5. Conclusion

The Affect-in-the-Wild (Aff-W) Challenge targets at bench-marking the efforts made in the research field of facial affect analysis in-the-wild. To develop the Aff-wild benchmark we collected and annotated the first large scale 'in-the- wild' database of facial affect, consisting of more

than 30 hours of videos showing reactions of about 200 persons, both males and females. In this paper, we described how the data have been collected and annotated. We give statistics of the benchmark. Furthermore, we describe the baseline system that we developed for the challenge. The performance of the baseline system demonstrates that the data are very challenging. Hence, they require meticulously designed deep learning approaches to be designed and implemented for the task. The different approaches which have been submitted to this challenge show that it is possible to develop new methods and obtain performance much higher than the baseline. Our current efforts are concentrated to obtain and annotate more videos, as well as annotate the previous videos with more attributes (e.g., facial action units).

6. Acknowledgments

The work of Stefanos Zafeiriou has been partially funded by the FiDiPro program of Tekes (project number: 1849/31/2015). The work of Dimitris Kollias was funded by a Teaching Fellowship of Imperial College London. We would like also to acknowledge the contribution of the Youtube users that gave us the permission to use their videos (especially Zalzar and Eddie from The1stTake).

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] J. Alabort-i-Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 679–682, New York, NY, USA, 2014. ACM.

[3] S. Albanie and A. Vedaldi. Learning grimaces by watching tv. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[4] M. S. Aung, S. Kaltwang, B. Romera-paredes, B. Martinez, A. Singh, M. Cella, M. F. Valstar, H. Meng, A. Kemp, A. C. Elkins, N. Tyler, P. J. Watson, A. C. Williams, M. Pantic, and N. Berthouze. The automatic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset. *IEEE Transactions on Affective Computing*, 2016.

[5] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 223–230. IEEE, 2006.

[6] W.-Y. Chang, S.-H. Hsu, and J.-H. Chien. Fatauva-net : An integrated deep learning framework for facial attribute recognition, action unit (au) detection, and valence-arousal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017.

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[8] G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou. A comprehensive performance evaluation of deformable face tracking" in-the-wild". *arXiv preprint arXiv:1603.06015*, 2016.

[9] C. Corneanu, M. Oliu, J. Cohn, and S. Escalera. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[10] R. Cowie and R. R. Cornelius. Describing the emotional states that are expressed in speech. *Speech communication*, 40(1):5–32, 2003.

[11] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder. 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[12] R. Cowie, G. McKeown, and E. Douglas-Cowie. Tracing emotion: an overview. *International Journal of Synthetic Emotions (IJSE)*, 3(1):1–17, 2012.

[13] T. Dalgleish and M. Power. *Handbook of cognition and emotion*. John Wiley & Sons, 2000.

[14] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 461–466. ACM, 2014.

[15] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516. ACM, 2013.

[16] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426. ACM, 2015.

[17] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[18] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2983–2991, 2015.

[19] D. Kollias, M. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[21] Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 253–256. IEEE, 2010.

[22] A. Lee. Welcome to virtualdub. org!-virtualdub. org, 2002.

[23] J. Li, Y. Chen, S. Xiao, J. Zhao, S. Roy, J. Feng, S. Yan, and T. Sim. Estimation of affective level in the wild with multiple memory networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017.

[24] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[25] M. Mahoor and B. Hasani. Facial affect estimation in the wild using deep residual and convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017.

[26] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014.

[27] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, 2012.

[28] M. A. Nicolaou, S. Zafeiriou, and M. Pantic. Correlated-spaces regression for learning continuous emotion dimensions. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 773–776. ACM, 2013.

[29] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.

[30] R. Plutchik. *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division, 1980.

[31] J. A. Russell. Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology*, 36(10):1152, 1978.

[32] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(6):1113–1133, 2015.

[33] Y.-l. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115, 2001.

[34] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2016.

[35] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65, 2010.

[36] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014.

[37] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013.

[38] A. Vedaldi and K. Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM, 2015.

[39] C. Whissel. The dictionary of affect in language, emotion: Theory, research and experience: vol. 4, the measurement of emotions, r. *Plutchik and H. Kellerman, Eds., New York: Academic*, 1989.

[40] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference On*, pages 1–6. IEEE, 2008.

[41] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006.

[42] L. YouTube. Youtube. *Retrieved*, 27:2011, 2011.

[43] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao. Facial affect"in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–47, 2016.

[44] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.